Published in final edited form as:
Comput Stat Data Anal. 2013 January 1; 56(1): . doi:10.1016/j.csda.2011.07.004.

A composite likelihood approach for spatially correlated survival data

Jane Paik[a,*] and **Zhiliang Ying**[b]

[a]Department of Medicine, Stanford University, Stanford, CA 94305, United States

[b]Department of Statistics, Columbia University, New York, NY 10025, United States

Abstract

The aim of this paper is to provide a composite likelihood approach to handle spatially correlated survival data using pairwise joint distributions. With e-commerce data, a recent question of interest in marketing research has been to describe spatially clustered purchasing behavior and to assess whether geographic distance is the appropriate metric to describe purchasing dependence. We present a model for the dependence structure of time-to-event data subject to spatial dependence to characterize purchasing behavior from the motivating example from e-commerce data. We assume the Farlie-Gumbel-Morgenstern (FGM) distribution and then model the dependence parameter as a function of geographic and demographic pairwise distances. For estimation of the dependence parameters, we present pairwise composite likelihood equations. We prove that the resulting estimators exhibit key properties of consistency and asymptotic normality under certain regularity conditions in the increasing-domain framework of spatial asymptotic theory.

Keywords

Spatial dependence; Pairwise joint likelihood; Marginal likelihood; Event times; Consistency; Asymptotic normality; Censoring

1. Introduction

Multivariate time-to-event data are subject to spatial correlation in familial and multicenter clinical trials in biomedical sciences, region-wide disease studies in epidemiology and e-commerce studies in marketing (Li and Lin, 2006; Henderson et al., 2002; Li and Ryan, 2002; Banerjee et al., 2003). The practical interest lies in the dependence between the survival outcomes in the geographic domain of interest. In standard geo-statistical practice, the variance and correlation structure of uncensored data are modeled through a parametric covariance function and the parameters are estimated by maximum likelihood (Cressie, 1993). The estimation of parameters poses a challenge, since using a full likelihood approach is computationally burdensome due to high-dimensional integrals. An added challenge in the analysis of time-to-event data prone to spatial correlation is the presence of censoring.

*Corresponding author. janepaik@stanford.edu (J. Paik).

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.07.004.

Two widely used methods in modeling associations between failure times are frailty and copula models. Nielsen et al. (1992), Klein (1992), Murphy (1995, 1996), and Parner (1998) investigated the estimation and inference for the frailty model. A comprehensive review can be found in Andersen et al. (1993) and Hougaard (2000). The frailty model-based approach is appealing for family studies since it accommodates the dependencies among relatives by assuming a shared frailty (Parner, 1998; Bandeen-Roche and Liang, 1996; Hsu and Gorfine, 2006). Models developed for clustered data may not fully allow for spatially correlated data. Motivated by a study of asthma onset in Boston, Li and Ryan (2002) used an extended frailty model to take into account the spatial correlation structure. Henderson et al. (2002) investigated survival of leukemia in northwest England by using a multivariate gamma frailty model with a covariance structure allowing for spatial effects.

Genest and MacKay (1986), Oakes (1989) and Shih and Louis (1995) modeled the association of bivariate failure times using copula functions. An attractive feature of the copula model is that the margins do not depend on the choice of the dependency structure. As a result, one can model and estimate the dependency and margins separately. In a multivariate setting, Li and Lin (2006) developed a method for analyzing survival data correlated in a region by specifying a multivariate normal copula and allowing for a spatial correlation structure in the parameters. They provided an estimating equation approach, avoiding the full likelihood that can be intractable when spatially correlated survival outcomes are involved.

Composite likelihood as proposed in Lindsay (1998) is convenient in the setting where the full likelihood is difficult to construct. Earlier, Besag (1974) considered a similar approach for spatial data. Cox and Reid (2004) provided a general framework for the composite likelihood approach to inference. Composite likelihood methods have been used in multivariate analysis of various types such as non-normal spatial data (Heagerty and Lele, 1998; Varin et al., 2005) and binary correlated data (LeCessie and Van Houwelingen, 1994; Kuk and Nott, 2000). Kuk (2007) considered a weighted composite likelihood for clustered data. Varin (2008) provided a survey of composite likelihood applications.

For survival data, the composite likelihood approach has been used in the analysis of clustered data in familial studies. Parner (2001) modeled the marginal distribution of pairs of failure times using shared frailty models and constructed a pseudo log-likelihood function by adding the pairwise likelihood contributions. Andersen (2004) specified joint survivor functions with copula models and estimated the marginal hazard and association parameters via composite likelihood. Tibaldi et al. (2004a,b) considered composite marginal likelihood inference for multivariate survival data using a Plackett–Dale (Plackett, 1965) model. For estimation, Zhao and Joe (2005) considered a two-stage approach and proposed the use in a multivariate setting in frailty and copula models. These methods are not suited for spatial correlation among the observations since they assume clusters to be independent.

The methodology in this paper is motivated by a problem in e-commerce data, where marketing research has generated much interest in ascertaining whether there is any spatial clustering in purchasing behavior among new customers (Bell and Song, 2007; Bradlow et al., 2005). We present a model for the dependence structure of failure times subject to spatial correlation. We use the Farlie–Gumbel–Morgenstern bivariate family as the survivor function. In order to address the question whether the spatial clustering of purchasing behavior depends on only geographical distances, we model the dependence parameter as a function of Euclidean distances and other pairwise distances. To estimate parameters, we follow a composite likelihood approach for the analysis of spatially correlated survival data. Following Cox and Reid (2004), we use a univariate marginal likelihood to estimate parameters in the hazard function and pairwise composite likelihood for the dependence

parameters. The resulting pairwise composite likelihood has a convenient form. We consider the case in which the marginal distribution of failure times follows the parametric Weibull family or the Cox proportional hazards model. We use a two-stage estimation procedure to estimate parameters from the marginal likelihood and then use composite likelihood to estimate dependence parameters. We prove that the resulting estimators exhibit key properties of consistency and asymptotic normality under certain regularity conditions in the increasing-domain spatial asymptotic framework.

In Section 2 we present the model and we present the estimation procedure of composite likelihood along with the asymptotic properties of the estimators. In Section 3 we discuss an application to a marketing study of e-commerce data and conclude with final remarks in Section 4.

2. Method

2.1. Notation and model

In a spatial region of interest, consider a total of n subjects who are followed up to failure or censoring. Let T be the failure time and C the censoring time. Let Z be a p-vector of covariates. Conditional on Z, T and C are assumed to be independent. Let $(\tilde{T}_i, C_i, Z_i, i = 1, \ldots, n)$ be n copies of (T, C, Z). For the ith subject, one can only observe $(\tilde{T}_i, Z_i, \delta_i, i = 1,\ldots, n)$ where $\tilde{T}_i = min(T_i, C_i)$ and $\delta_i = 1$ $(T_i \le C_i)$. Denote by $\lambda_0(t)$ the baseline hazard function and $\Lambda_0(t)=\int_0^t \lambda_0(s)\mathrm{d}s$ the baseline cumulative hazard function. Let $\Lambda_i(t)$ denote the cumulative hazard function, and let β be a p-dimensional regression coefficient vector. In addition to the covariates, each subject's geographic coordinates (latitude and longitude) are observed. We are interested in the marginal effect of the covariate vector Z on the hazard function as well as the dependence structure among failure times.

A first step towards modeling the dependence structure is to model pairwise joint distributions. We propose using the Farlie-Gumbel-Morgenstern (FGM) family of distributions (Morgenstern, 1956; Gumbel, 1960; Farlie, 1960), given by

$$F_{ij}(t_i, t_j)=F_i(t_i)F_j(t_j)[1+\xi_{ij}\{1 - F_i(t_i)\}\{1 - F_j(t_j)\}],$$

where $F_i(t_i)$ is the marginal survival function of T_i, and $F_{ij}(t_i, t_j) = P(T_i \ge t_i, T_j \ge t_j)$. There is a restriction on ξ_{ij} such that $0 \le |\xi_{ij}| \le 1$. The parameter ξ_{ij} can be viewed as a measure of dependence. We consider the case in which the marginals follow the Weibull family $F_i(t_i)=\exp[-t_i^{\gamma}\exp(\alpha+\beta' Z_i)]$.

To parameterize the bivariate joint distribution, it suffices to parameterize ξ_{ij} and the univariate marginal survival functions $F_i(t_i)$. To this end, we propose that the dependence parameter ξ_{ij} is itself a function of geographic distance as well as other arbitrary pairwise distances. Specifically, let $\xi_{ij} = \xi_{ij}(\psi; w_{ij})$, where w_{ij} could include the Euclidean distance and certain pairwise characteristics and ψ is a vector of parameters to be estimated.

Specifically, let d_{ij} be the Euclidean distance between the spatial locations of subjects i and j, where $d_{ii} = 0$ by definition, and z_{ij} is a function of demographic variables for units i and j, e.g. an indicator of whether subjects i and j reside in metropolitan regions.

$$\xi_{ij}(\psi;w_{ij})=cv_{ij}\exp(d_{ij}\psi_1),$$

where

$$v_{ij}=\frac{\exp(\psi_0+z_{ij}\psi_2)}{1+\exp(\psi_0+z_{ij}\psi_2)},$$

and c is a known constant, which will be shown below. The exponential decay $\exp(d_{ij}, \psi)$ is modified by a factor of v_{ij} which varies depending on pairwise characteristics. This parameterization allows for the spatial dependence to vary among observations of varying distances as well as varying demographic profiles.

It is clear that the case of $\xi_{ij} = 0$ corresponds to case where T_i and T_j are independent. To better understand the meaning of ξ_{ij} for non-zero values of ξ_{ij}, we show that the dependence parameter ξ_{ij} is related to a global measure of dependence discussed by Hsu and Prentice (1996). The global measure is defined as $\rho_{ij} := corr\{\Lambda_i(T_i), \Lambda_j(T_j)\}$, $i \neq j = 1,\ldots, n$. We show that the global measure of dependence is in one-to-one correspondence with the parameter ξ_{ij}. Hsu and Prentice (1996) present the following relationship:

$$\rho_{ij}=\int_0^\infty\int_0^\infty \Lambda_i(T_i)\Lambda_j(T_j)F(dT_i, dT_j) - 1.$$

Using the notation where $F_i(t_i) = F_{i0}\{\Lambda_i(t_i)\}$ and $F(t_i, t_j) = F_0\{\Lambda_i(t_i), \Lambda_j(t_j)\}$, we can show that for a given pair, the correlation between subject i and j is given by

$$\begin{aligned}\rho_{ij}&=\int\int F_0\{x_i, x_j\}dx_i dx_j - 1\\ &=\int\int F_{i0}(x_i)F_{j0}(x_j)[1+\xi\{1 - F_{i0}(x_i)\}\{1 - F_{j0}(x_j)\}]dx_i dx_j - 1\\ &=\xi\int_0^\infty F_{i0}(x_i)\{1 - F_{i0}(x_i)\}dx_i\int_0^\infty F_{j0}(x_j)\{1 - F_{j0}(x_j)\}dx_j.\end{aligned}$$

This implies that the global dependence measure, ρ_{ij}, is related to ξ_{ij} by a constant. If we specify $F_{i0}(x_i) = e^{-x_i}$ and $F_{j0}(x_j) = e^{-x_j}$, then $\rho_{ij} = \xi_{ij}/4$ and $c = 4$.

2.2. Inferential procedure and asymptotic properties

Let β_0 be a vector of p regression coefficients and let $\lambda(t|z)$ be the marginal hazard rate of an individual with covariate value z. We consider the case $\lambda(t|Z) = \lambda_0(t)\exp(\beta'Z)$. Using the usual counting process notation, let $N_i(t) = 1(\tilde{T_i} \leq t, \delta_i = 1)$ be the counting process which jumps at time T_i if $T_i \leq C_i$, and the at-risk process $Y_i(t) = 1(\tilde{T_i} \geq t)$. Let $M_i(t)$ be the corresponding martingale adapted to the filtration $F_{i,t} = \sigma\{N_i(s), Y_i(s), Z_i, 0 \leq s < t\}$,

$$M_i(t)=N_i(t)-\int_0^t Y_i(s)\exp(\beta' Z_i)d\Lambda_0(s).$$

Suppose that the baseline hazard $\lambda_0(t)$ is specified up to a finite dimensional parameter η, i.e. $\lambda_0(t) = \lambda_0(t, \eta)$. Let $\theta = (\beta, \eta')'$. Our strategy is to construct a joint estimating equation to estimate θ via a marginal likelihood, and estimate ψ via the composite likelihood shown below. Let $F_i(t, \theta) = \exp\{-\exp(\beta'Z_i)\Lambda_0(t, \eta)\}$. Then the log-likelihood function has the form

$$l_M(\beta,\eta)=\sum_{i=1}^n[\delta_i\{\beta' Z_i+\log\lambda_0(\tilde{T}_i,\eta)\}+\log F_i(\tilde{T}_i,\theta)].$$

Specifying the marginal model as a Weibull model, we have $\lambda_0(t, \eta) = \gamma t^{\gamma-1}\exp(\alpha)$, where $\eta = (\alpha, \gamma)'$. The log-likelihood has the following form

$$l_M(\theta)=\sum_{i=1}^{n}\left[(\gamma-1)\delta_i\log\tilde{T}_i+\delta_i(\log\gamma+\alpha+\beta' Z_i)-\tilde{T}_i^{\gamma}\exp(\alpha+\beta' Z_i)\right].$$

Thus we have the score function $\sum_{i=1}^{n} U_i(\theta)$ where

$$U_i(\theta)=\begin{pmatrix} Z_i\{\delta_i-\tilde{T}_i^{\gamma}\exp(\alpha+\beta' Z_i)\} \\ \{\delta_i-\tilde{T}_i^{\gamma}\exp(\alpha+\beta' Z_i)\} \\ \delta_i(\log\tilde{T}_i+1/\gamma)-\tilde{T}_i^{\gamma}\log\tilde{T}_i\exp(\alpha+\beta' Z_i) \end{pmatrix}.$$

We will use pairwise composite likelihood to estimate the spatial dependence parameter ψ. There exists a multivariate distribution whose bivariate survivor functions are $F_{ij}(t_i, t_j, \theta, \psi)$ (Prentice and Cai, 1992). Below we show an estimating function for ψ when the marginal hazard is known.

The pairwise composite likelihood has the following general expression:

$$L_c(\psi,\theta)=\prod_{\substack{i,j\in D_n\\ i\le j}} F_{ij}(\tilde{T}_i,\tilde{T}_j;\theta,\psi)^{(1-\delta_i)(1-\delta_j)}\{F_{ij}^{(1)}(\tilde{T}_i,\tilde{T}_j;\theta,\psi)\}^{\delta_i(1-\delta_j)}\{F_{ij}^{(2)}(\tilde{T}_i,\tilde{T}_j;\theta,\psi)\}^{(1-\delta_i)\delta_j}F_{ij}(\tilde{T}_i,\tilde{T}_j;\theta,\psi)^{\delta_i\delta_j}$$

where

$$\begin{aligned} F_{ij}^{(1)}(t_i,t_j;\theta,\psi)&=-\tfrac{\partial}{\partial t_i}F_{ij}(t_i,t_j;\theta,\psi)\\ &=f_i(t_i;\theta)F_j(t_j;\theta)[1+\xi_{ij}(\psi)\{1-2F_i(t_i;\theta)\}\{1-F_j(t_j;\theta)\}],\\ F_{ij}^{(2)}(t_i,t_j;\theta,\psi)&=-\tfrac{\partial}{\partial t_j}F_{ij}(t_i,t_j;\theta,\psi)\\ &=F_i(t_i;\theta)f_j(t_j;\theta)[1+\xi_{ij}(\psi)\{1-F_i(t_i;\theta)\}\{1-2F_j(t_j;\theta)\}], \end{aligned}$$

and the joint density of T_i and T_j is given by

$$f_{ij}(t_i,t_j;\theta,\psi)=f_i(t_i;\theta)f_j(t_j;\theta)[1+\xi_{ij}(\psi)\{1-2F_i(t_i;\theta)\}\{1-2F_j(t_j;\theta)\}].$$

We note that the summation involves all pairs i and j that belong to D_n, a sequence of strictly increasing finite domains of Z^d with cardinality $|D_n|$. The summand is a $|D_n|$-dimensional vector random variable. Then

$$\begin{aligned} \ell_c(\psi,\theta)=\sum_{\substack{i,j\in D_n\\ i\le j}}\;&[(1-\delta_i)(1-\delta_j)\log[1+\xi_{ij}(\psi)\{1-F_i(\tilde{T}_i,\theta)\}\{1-F_j(\tilde{T}_j;\theta)\}]\\ &+\delta_i(1-\delta_j)\log\{1+\xi_{ij}(\psi)\{1-2F_i(\tilde{T}_i;\theta)\}\{1-F_j(\tilde{T}_i;\theta)\}\}\\ &+\delta_j(1-\delta_i)\log\{1+\xi_{ij}(\psi)\{1-F_i(\tilde{T}_i;\theta)\}\{1-2F_j(\tilde{T}_i;\theta)\}\}\\ &+\delta_i\delta_j\log\{1+\xi_{ij}(\psi)\{1-2F_i(\tilde{T}_i;\theta)\}\{1-2F_j(\tilde{T}_i;\theta)\}\}]+c \end{aligned}$$

where c denotes the terms that depend on the marginal distribution only and do not involve the parameter ψ for the dependence. The pairwise score equation is

$$\sum_{\substack{i,j \in D_n \\ i \le j}} U_{ij}(\psi,\theta) = \sum_{\substack{i,j \in D_n \\ i \le j}} = \left[\frac{(1-\delta_i)(1-\delta_j)\{1-F_i(\tilde{T}_i,\theta)\}\{1-F_j(\tilde{T}_i;\theta)\}}{1+\xi_{ij}(\psi)\{1-F_i(\tilde{T}_i,\theta)\}\{1-F_j(\tilde{T}_i;\theta)\}} + \frac{\delta_i(1-\delta_j)\{1-2F_i(\tilde{T}_i;\theta)\}\{1-F_j(\tilde{T}_j;\theta)\}}{1+\xi_{ij}(\psi)\{1-2F_i(\tilde{T}_i;\theta)\}\{1-F_j(\tilde{T}_i;\theta)\}} +$$

Where $\dot{\xi}_{ij}(\psi) = \frac{\partial}{\partial \xi}\xi_{ij}(\psi)$. We adopt a two-stage approach for estimation. In the first stage, we apply the estimating procedure for the Weibull model to obtain estimates of regression parameters and the baseline hazard. Estimates from the first stage are plugged into the estimating Equation 1 to solve for the spatial parameters $\hat{\psi}$. This two-stage approach can be expressed as a joint estimating equation of the marginal and dependence parameter, since the estimating equation for the marginal hazards does not involve the dependence parameter. We solve the following equation

$$\begin{pmatrix} \sum_{i=1}^{n} U_i(\theta) \\ \sum_{\substack{i,j \in D_n \\ i \le j}} U_{ij}(\psi,\theta) \end{pmatrix} = 0.$$

Let $\hat{\theta}$ and $\hat{\psi}$ denote the solutions to the joint equation above. We present the asymptotic result for the Weibull model.

Theorem 3.1—Under suitable regularity conditions,

$$\sqrt{n}\begin{pmatrix} \hat{\theta} & \theta_0 \\ \hat{\psi} & \psi_0 \end{pmatrix} \to N(0, \sum)$$

where

$$\Sigma = \begin{pmatrix} A_{11.2}^{-1} & 0 \\ A_{22.1}^{-1}A_{21}A_{11.2}^{-1} & A_{22.1}^{-1} \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} A_{11.2}^{-1} & 0 \\ A_{22.1}^{-1}A_{21}A_{11.2}^{-1} & A_{22.1}^{-1} \end{pmatrix}'$$

$$A_{11} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial U_i(\theta)}{\partial \theta'},$$

$$A_{21} = \lim_{n\to\infty} \frac{1}{n} \sum_{\substack{i \in D_n, d(i,j) \le r_m \\ j \in D_n, d(i,j) \le k_m}} \frac{\partial U_{ij}(\psi,\theta)}{\partial \theta'},$$

$$A_{22} = \lim_{n\to\infty} \frac{1}{n} \sum_{\substack{i \in D_n, d(i,j) \le r_m \\ j \in D_n, d(i,j) \le k_m}} \frac{\partial U_{ij}(\psi,\theta)}{\partial \psi'},$$

$$V_{11} = \lim_{n\to\infty} \frac{1}{n} \sum_{\substack{i,j \in D_n \\ d(i,j) \le k_m}} U_i(\theta) U_j(\theta)',$$

$$V_{12} = \lim_{n\to\infty} \frac{1}{n} \sum_{\substack{i \in D_n, d(i,j) \le r_m \\ j,k \in D_n \\ d(i,j) \text{ or } d(j,k) \le k_m}} U_k(\theta) U_{ij}(\psi,\theta)',$$

$$V_{22} = \lim_{n\to\infty} \frac{1}{n} \sum_{\substack{i,j,k,l \in D_n \\ d(i,j) \le r_m \\ d(j,k) \text{ or } d(k,l) \le k_m}} U_{ij}(\psi,\theta) U_{kl}(\psi,\theta)',$$

$$A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}, \quad \text{and}$$

$$A_{22.1}^{-1} = A_{22}^{-1} + A_{22}^{-1}A_{21}A_{11.2}^{-1}A_{12}A_{22}^{-1}.$$

We can consistently estimate A_{ij} and V_{ij} by replacing θ and ψ in $U_i(\theta)$, $U_{ij}(\psi, \theta)$ by $\hat{\theta}$ and $\hat{\psi}$. To ensure consistent variance estimation, we assume a sequence of strictly increasing finite domains D_n of Z^d that satisfy the following condition: there exists $\alpha > 0$ and (n_m) a strictly increasing sequence of integers such that

$$\sum_{m\ge 1} m^{\alpha} |D_{n_m}|^{-1} < \infty$$

$$\sum_{m\ge 1} m^{\alpha} \left(\frac{|D_{n_{m+1}} - D_{n_m}|}{|D_{n_m}|} \right)^2 < \infty.$$

We choose a sequence $\{k_m\}$ satisfying $k_m^d = O(m^{\beta})$ with $0 < \beta < \alpha/2$, and also choose $r_m \to \infty$ such that $r_m < \frac{1}{3} k_m$. For consistent variance estimation, the neighborhoods are chosen by the distance between i and j denoted as $d(i, j)$.

The pairwise score functions can essentially be formulated as weighted U-statistics as shown in Equation (4) on page 2 of a separate Technical Report available from the authors. Since none of the existing spatial asymptotic theory works for U-statistics, the supplementary material provides relevant theory for U-statistics considered in the spatial domain. When the marginal hazards model is assumed to follow a Weibull model, we can invoke Theorem 1 from the Technical Report for the asymptotic normality of the estimator based on the marginal likelihood and Theorem 2 obtain the asymptotic normality of the estimator of the spatial parameter based on the pairwise composite likelihood. We appeal to Theorem 5 to

justify the consistency of the variance estimator derived from the marginal likelihood by showing the difference between the variance estimator and theoretical quantity converges to 0 in probability. Under suitable mixing conditions, Theorem 6 shows that consistent variance estimators based on the composite pairwise likelihood can be obtained.

We can replace the Weibull model with the Cox proportional hazards model and develop the inferential procedure similarly. In this case, the composite likelihood involves the unknown baseline hazard and one would proceed by plugging in the Aalen-Breslow type estimator. For the Cox model, we report asymptotic results in Web Appendix I.

3. Marketing data

We apply the proposed method to data from a marketing study on e-commerce. In the marketing literature, there is a significant interest in finding the appropriate metric to describe spatial dependence; more specifically, the interesting question is to determine whether spatial dependence is solely due to geographic distance. The data source is an internet retailer that has documented the calendar time of a business launch and the purchases made in a time period. Demographic and local retail information on the zip codes comes from a secondary data source (ZIP Business Patterns database from the US Census Bureau). The research interest is to test whether a neighborhood effect is present among e-commerce related to the internet retailer and to determine the factors affecting internet retailing. We consider each observation as the time from the start of the business until the first purchase in the zip code. If there was no purchase in the zip code, the observation was censored. The data include 1596 zip codes in the state of New York.

Figs. 1 and 2 display the region of interest. The initial time point is the launch of the website, and the times until purchase are plotted at the geographical coordinates of each customer. Circles are plotted inversely proportional to the length of time until purchase.

We fit a parametric Weibull model and a Cox model for the marginal hazard. Since the size of a neighborhood may be related to the hazard of the first purchase, we included the number of supermarkets in each zip code as a covariate to account for varying population sizes. Based on prior marketing research, the following covariates are selected: the number of supermarkets in the zip code, percentage of households with children, percentage of households who earn greater than $50,000 annually, and the percentage of households with at least one family member who attended college. For the percentage of households with at least one household member with a college degree, the Weibull model yielded a log hazard ratio of 2.629, and the Cox proportional hazards model yielded an estimate of 2.976. This indicates neighborhoods with higher proportions of educated households are more likely to purchase from the web-based company. A summary of the marginal results from the Weibull models and Cox proportional hazards are combined in Table 1.

The model for dependence attributes pairwise characteristics of the population of interest to the dependence of purchasing times. In the marketing literature, this dependence is referred to as a neighborhood effect (Bell and Song, 2007; Bradlow et al, 2005). We allowed the dependence parameter to be a function of geographic distance and an indicator characterizing the type of residential neighborhood of the pairs. Tables 1 and 2 show a summary of parameter estimates of the dependence model for the Weibull and Cox models. The estimates correspond to the intercept, geographic distance between the pair, and an indicator for the given pair residing in metropolitan areas. Both marginal hazards models yielded virtually identical estimates of ψ up to 4 decimal points. While the updating values were different, the two algorithms converged in 7 steps.

The estimate corresponding to Euclidean distance (−0.603, standard error 0.068) indicates that the dependence decreases as geographic distance increases. For example, the dependence ξ between pairs living in metropolitan regions is 0.107 when the distance between the pair is 0.5 km, and is 0.079 when the distance between the pair is 1 km. When the distance is 2 km, the dependence is 0.044. This confirms a previous finding in the marketing science literature that the correlation between buyers decreases as a function of geographic distance (Bell and Song, 2007). We note that testing $\psi_2 = 0$ corresponds to testing for the equality of dependence among metropolitan pairs versus non-metropolitan pairs given the same distance. The estimate for ψ_2 indicates that for a given distance, the dependence between two people of metropolitan areas is larger than it is between two people who reside in non-metropolitan regions. When either member of the pair resides in non-metropolitan regions, the dependence among the pair is 0.03 when the distance is 1 km apart.

A significant interest in the marketing literature is to assess whether geographic distance is the appropriate metric to describe purchasing dependence. Our findings add to the results in the literature that the spatial dependence of e-commerce may be more pronounced in urban than rural areas. The interpretation of the model estimates is that residents of different neighborhood types have varying degrees of spatial dependence.

4. Discussion

In this paper, we proposed a composite likelihood approach for estimating parameters in a semiparametric model for spatially correlated survival data. In the marketing data example, the goals were to determine factors affecting internet purchasing patterns in a population of interest and to characterize the dependence of purchasing behavior. Our model is suitable for the goals of this study. Frailty models such as those proposed in Banerjee et al. (2003) or Henderson et al. (2002) assume a subject-specific frailty, and parameter estimates do not carry an interpretation of a population average. Moreover, these models may not be well suited to address the question of interest as it would be difficult to model the dependence in as a function of anything other than distance. On the other hand, another marginal approach has been proposed by Li and Lin (2006). Their model makes use of a multivariate normal copula which has an estimation procedure that is computationally complex due to numerical integration. Our approach does not require numerical integration but follows a straightforward two-step procedure. This two-stage approach can be expressed as a joint estimating equation of the marginal and dependence parameter, since the estimating equation for the marginal hazards does not involve the dependence parameter. Further, we note that the estimating equation for β is unbiased even when the spatial correlation structure is misspecified. A limitation is that the FGM family may be only able to handle small dependence, although in our marketing example the dependence among purchasing behavior is moderately low.

In this paper, we considered the FGM family of bivariate distributions through the specification of marginals, but the FGM family is found to be a special case of a broader class of models using a representation proposed by Prentice and Cai (1992).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank David Bell of the Wharton School of the University of Pennsylvania for providing the data on e-commerce as well as helpful discussions.

References

Andersen, PJ.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. Springer; New York: 1993.

Andersen E. Composite likelihood and two-stage estimation in family studies. Biostatistics. 2004; 5(1):15–30. [PubMed: 14744825]

Bandeen-Roche KJ, Liang KY. Modeling failure-time associations in data with multiple levels of clustering. Biometrika. 1996; 83:29–39.

Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics. 2003; 4:123–142. [PubMed: 12925334]

Bell DR, Song S. Neighborhood effects and trial on the internet: Evidence from online grocery retailing. Quant Marketing Econom. 2007; 5:361–400.

Besag J. Spatial interaction and the statistical analysis of lattice systems (with Discussion). J R Statist Soc B. 1974; 36:192–236.

Bradlow ET, et al. Spatial Models in Marketing. Marketing Lett. 2005; 16:267–278.

Cox DR, Reid N. A Note on pseudolikelihood constructed from marginal densities. Biometrika. 2004; 91(3):729–737.

Cressie, N. Statistics for Spatial Data. Wiley; New York: 1993.

Farlie DJG. The performance of some correlation coefficients for a general bivariate distribution. Biometrika. 1960; 47:307–323.

Genest C, MacKay RJ. Copules archimédiennes et familles de lois bidimensionelles dont les marges sont données. Canad J Statist. 1986; 14:145–159.

Gumbel EJ. On Bivariate exponential distributions. J Amer Statist Assoc. 1960; 55:698–707.

Heagerty PJ, Lele SR. A Composite likelihood approach to binary spatial data. J Amer Statist Assoc. 1998; 93:1099–1111.

Henderson R, Shimakura S, Gorst D. Modeling spatial variation in leukemia survival data. J Amer Statist Assoc. 2002; 97:965–972.

Hougaard, P. Analysis of Multivariate Survival Data. Springer; New York: 2000.

Hsu L, Gorfine M. Multivariate survival analysis for case-control family data. Biostatistics. 2006; 7(3): 387–398. [PubMed: 16368774]

Hsu L, Prentice R. On Assessing the strength of dependency between failure time variates. Biometrika. 1996; 83(3):491–506.

Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics. 1992; 48:795–806. [PubMed: 1420842]

Kuk AYC. A hybrid pairwise likelihood method. Biometrika. 2007; 94:939–952.

Kuk AYC, Nott DJ. A pairwise likelihood approach to analyzing correlated binary data. Statist Probab Lett. 2000; 47:329–335.

LeCessie S, Van Houwelingen JC. Logistic regression for correlated binary data. Appl Statist. 1994; 43:95–108.

Li Y, Lin X. Semiparametric normal transformation models for spatially correlated survival data. J Amer Statist Assoc. 2006; 101(474):591–603.

Li Y, Ryan L. Modeling spatial survival data using semi-parametric frailty models. Biometrics. 2002; 58:287–297. [PubMed: 12071401]

Lindsay BG. Composite likelihood methods. Contemp Math. 1998; 80:221–239.

Morgenstern D. Einfache Beispiele zweidimensionaler Verteilungen. Mitteilungsblatt für Mathematische Statistik. 1956; 8:234–235.

Murphy SA. Consistency in a proportional hazards model incorporating a random effect. Ann Statist. 1995; 22:712–731.

Murphy SA. Asymptotic theory for the frailty model. Amer Statist. 1996; 23:183–214.

Nielsen G, Andersen P, Gill R, Sorensen T. A Counting process approach to maximum likelihood estimation in frailty models. Scand J Stat. 1992; 19:25–43.

Oakes D. Bivariate survival models induced by frailties. J Amer Statist Assoc. 1989; 101(84):487–493.

Parner E. Asymptotic theory for the correlated gamma-frailty model. Ann Statist. 1998; 26:183–214.

Parner ET. A Composite likelihood approach to multivariate survival data. Scand J Stat. 2001; 28:295–302.

Plackett RL. A class of bivariate distributions. J Amer Statist Assoc. 1965; 60(310):516–522.

Prentice RL, Cai L. Covariance and survivor function estimation using censored multivariate failure time data. Biometrika. 1992; 79:495–512.Correction. 1993; 80:7112.

Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. Biometrics. 1995; 51:1384–1399. [PubMed: 8589230]

Tibaldi F, Molenberghs G, Burzykowski T, Geys H. Pseudolikelihood estimation for a marginal multivariate survival model. Stat Med. 2004a; 23:947–963. [PubMed: 15027082]

Tibaldi F, Barbosa FT, Molenberghs G. Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett–Dale model. Stat Med. 2004b; 23:2173–2186. [PubMed: 15236423]

Varin C. On composite marginal likelihoods. Adv Statist Anal. 2008; 92:1–28.

Varin C, Høst G, Skare Ø. Pairwise likelihood inference in spatial generalized linear mixed models. Comput Statist Data Anal. 2005; 49:1173–1191.

Zhao Y, Joe H. Composite likelihood estimation in multivariate data analysis. Canad J Statist. 2005; 33:335–356.

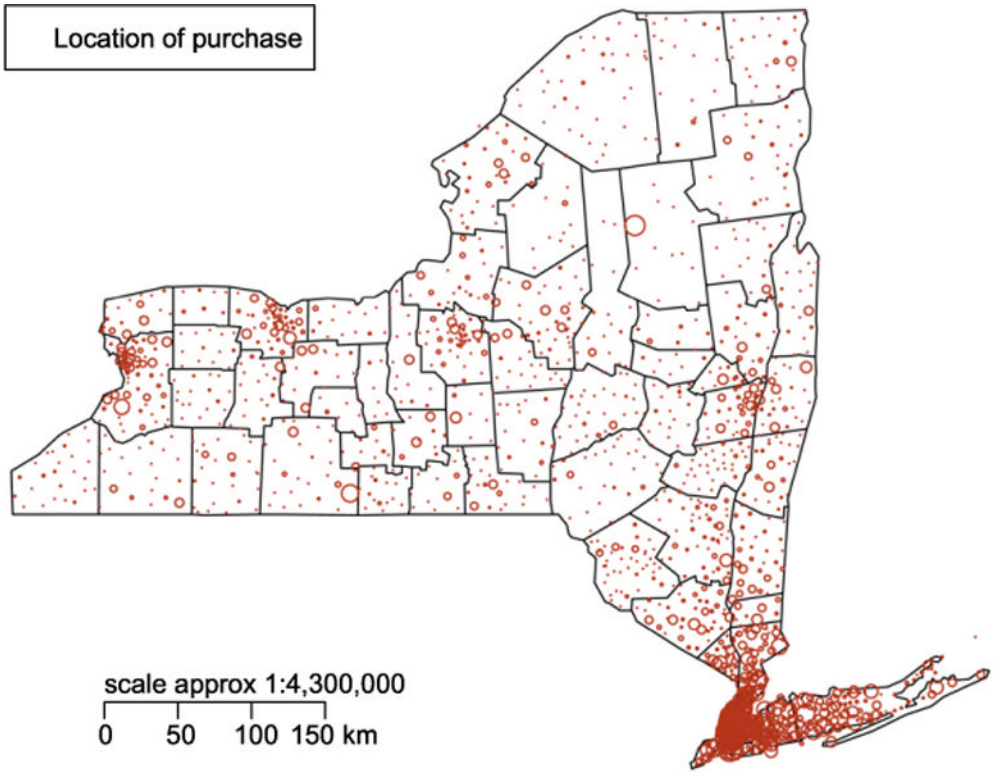

Fig. 1.
Customer purchases in New York State.

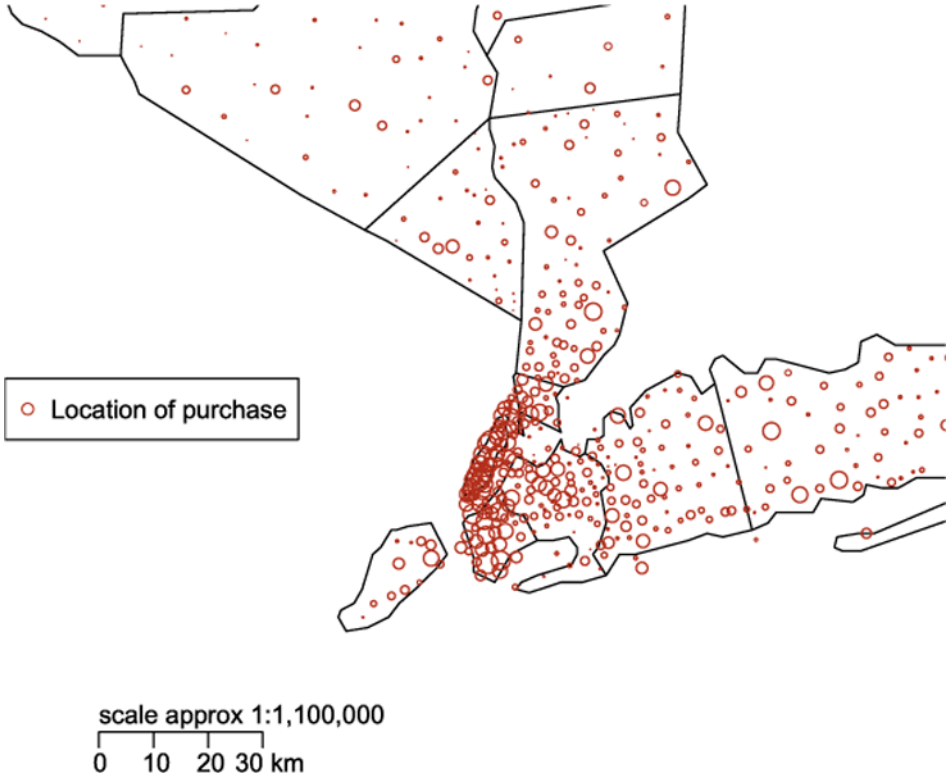

Fig. 2.
Customer purchases in a metropolitan region of New York.

Table 1

Marketing example: estimates for Marginal Hazard using Cox PH.

β	est	exp(β)	se	p
β_1	0.084	1.088	0.003	0.000
β_2	0.562	1.755	0.635	0.116
β_3	2.976	19.625	0.400	0.000
β_4	2.401	11.033	0.461	0.000

Table 2

Dependence parameter estimates.

Weibull			
ψ	**Parameter estimate**	**Std. error**	***p*-value**
ψ_0	0.339	0.014	<0.001
ψ_1	−0.603	0.068	<0.001
ψ_2	−1.656	0.057	<0.001